Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 161-164 recibido 23-03-2015 revisado 27-04-2015 aceptado 14-05-2015

IPHealth: Plataforma inteligente basada en *open*, *linked* y *big data* para la toma de decisiones y aprendizaje en el ámbito de la salud

IPHealth: Intelligent platform based on open, linked and big data for decision-making and learning in the field of health

Manuel de Buenaga
Diego Gachet
Dpto. de Sistemas Informáticos
Universidad Europea de Madrid
C/ Tajo s/n – 28670
Villaviciosa de Odón (Madrid)
buenaga@uem.es
diego.gachet@uem.es

Manuel J. Maña
Jacinto Mata
Dpto. de Tecnologías de la Información
Universidad de Huelva
Carretera Palos s/n – 21819
Palos de la Frontera (Huelva)
manuel.mana@dti.uhu.es
jacinto.mata@dti.uhu.es

L. Borrajo
E.L. Lorenzo
Dpto. de Informática
Universidad de Vigo
Edificio Politécnico
Campus Universitario
32004 Ourense
eva@uvigo.es
lborrajo@uvigo.es

Resumen: El proyecto IPHealth tiene como principal objetivo diseñar e implementar una plataforma con servicios que permitan un acceso integrado e inteligente a la información relacionada en el entorno biomédico. Se proponen tres escenarios de uso: (i) la asistencia a los profesionales sanitarios durante el proceso de toma de decisiones en el ámbito clínico, (ii) el acceso a información relevante sobre su estado de salud a pacientes crónicos y dependientes y (iii) el soporte a la formación basada en la evidencia de los nuevos estudiantes de medicina. Se propondrán técnicas más efectivas para diversas técnicas de PLN y extracción de información de grandes conjuntos de datos tanto provenientes de sensores como utilizando conjuntos de datos y texto de libre acceso. Se diseñará una arquitectura y un framework de aplicaciones Web que permita la integración de procesos y técnicas de minería de texto y datos e integración de información de una forma rápida, uniforme y reutilizable (mediante plugins).
Palabras clave: Minería de Texto, Minería de Datos, Integración de Información, Open Data, Big Data, Smart Sensors, Sistemas de Salud Personal

Abstract: The IPHealth project's main objective is to design and implement a platform with services that enable an integrated and intelligent access to related in the biomedical domain. We propose three usage scenarios: (i) assistance to healthcare professionals during the decision making process at clinical settings, (ii) access to relevant information about their health status and dependent chronic patients and (iii) to support evidence-based training of new medical students. Most effective techniques are proposed for reveral NLP tecniques and extraction of information from large data sets from sets of sensors and using open data. A Web application framework and an architecture that would enable integration of processes and techniques of text and data mining will be designed. Also, this architecture have to allow an integration of information in a fast, consistent and reusable (via plugins) way.
Keywords: Text Mining, Data Mining, Information Integration, Open Data, Big Data, Smart Sensors, Personal Health

1 Introducción

El proyecto IPHealth se centra en la aplicación de técnicas de integración de información para conseguir abarcar fuentes de tipo heterogéneo y de minería de textos y datos para facilitar la extracción de conocimiento útil en el contexto de la medicina personalizada (PM, Personalized Medicine). Es un proyecto financiado por el Ministerio de Economía y Competitividad, correspondientesa la convocatoria 2013 del Programa Estatal de Investigación, Desarrollo e

ISSN 1135-5948

Innovación Orientada a los Retos de la Sociedad (TIN2013-47153-C03) y que se desarrolla desde enero de 2014 hasta diciembre de 2016.

Los numerosos avances y descubrimientos en el ámbito biomédico que han tenido lugar en la última década, tanto a nivel de tecnología como de investigación básica, han supuesto un importante revulsivo en el enfoque y la práctica clínica moderna de la medicina basada en la evidencia (EBM, Evidence-Based Medicine) (Kumar, 2011) y la medicina personalizada.

La medicina personalizada busca la identificación de terapias personalizadas que hagan seguro y efectivo el tratamiento individualizado de pacientes específicos. Una de las grandes dificultades para llevar a cabo esta práctica clínica de forma efectiva es que en la actualidad no existen sistemas flexibles de información capaces de proporcionar conocimiento preciso, actualizado e interrelacionado basado en el acceso estratificado a múltiples orígenes de datos de tipo heterogéneo (Fernald, 2011). Toda esta información, generada en estudios experimentales, ensayos clínicos y en la práctica clínica diaria, así como recientemente a través de sensores biomédicos y grandes conjuntos de datos y texto de libre acceso y entrelazado (Open y Linked Data) debería convertirse en una fuente extraordinaria de conocimiento para el avance de la medicina personalizada. Sin embargo, la medicina personalizada se enfrenta en la actualidad a grandes retos. Es necesario integrar información heterogénea dispersa en múltiples orígenes, de diferentes género, dominio, estructura y escala, donde además juega un papel muy importante la componente textual.

2 Objetivos y beneficios del proyecto

El objetivo general del proyecto es analizar, experimentar y desarrollar nuevas técnicas de minería de texto e integración de información sobre grandes cantidades de datos de fuentes de información y conocimiento heterogéneas, como elementos clave en sistemas inteligentes de acceso a la información biomédica, desarrollando una plataforma que implemente este tipo de servicios. El objetivo general se desglosa en los siguientes objetivos concretos:

O1. Desarrollo de técnicas de minería de texto bilingües (Inglés y Español) adaptadas al dominio clínico, como generación de resúmenes, recuperación de imágenes a partir de texto, recuperación de información, reconocimiento de entidades nombradas y extracción de información.

O2. Diseño y desarrollo de una arquitectura Big-Data para el almacenamiento, análisis y consulta en tiempo real de datos de trabajo. Los datos de trabajo incluyen tanto los procedentes de sensores biomédicos inalámbricos como la obtenida de fuentes de datos públicas.

O3. Diseño de una plataforma Web para soporte e integración de técnicas inteligentes que permita a los usuarios un acceso remoto y amigable a las herramientas finales.

O4. Desarrollo de herramientas inteligentes para el soporte al usuario en la toma de decisiones para el diagnóstico y tratamiento, así como para la formación.

O5. Evaluación de la efectividad y la usabilidad a través de evaluaciones sistemáticas y con usuarios.

Los objetivos del proyecto se encuentran en sintonía con el Reto en Economía y Sociedad Digital del Plan Estatal de I+D, así como con el reto “Health, Demographic Change and Wellbeing” del programa Horizonte 2020.

3 Metodología

3.1 Arquitectura Big Data

Una importante contribución del proyecto ipHealth es el uso de Tecnologías Big Data tanto para el tratamiento como para la integración de información heterogénea en el campo de la salud, información que puede provenir de diversas fuentes como por ejemplo historiales clínicos que aportan texto o bien de sensores fisiológicos que proporcionan otro tipo de información. El diseño de una arquitectura hardware y software que permita el almacenamiento masivo de datos así como su procesamiento debe necesariamente organizarse en diferentes capas que permitan la captura de datos, su filtrado, almacenamiento y posterior análisis y consulta de resultados.

La arquitectura hardware/software que se plantea en el proyecto debe corresponderse con las necesidades planteadas en cuanto a capturar, almacenar y procesar grandes conjuntos de datos, del orden de cientos de GBytes, lo que implica el resolver problemas como la escalabilidad, de manera que podamos realizar una monitorización y análisis de datos eficiente para un gran número de pacientes o bien una gran cantidad de datos como los que pueden provenir de los sensores fisiológicos al mismo

tiempo que se mantiene un coste asumible. En este sentido hemos optado por utilizar desde el punto de vista de hardware una infraestructira de cloud mediante el servicio EC2 de Amazon (Pandeya, 2012), esta primera capa permite implementar las funcionalidades necesarias de captura y almacenamiento de la información, sobre ella una segunda capa con utilizamos herramientas y tecnología Hadoop, bases de datos no relacionales y lenguajes de programación como Java y R (Prajapati 2013) que nos permitan a su vez el desarrollo de algoritmos modernos, complejos y específicos de análisis y modelado de datos que permitan la obtención de conocimiento a partir de grandes cantidades masivas de información y que a su vez faciliten el descubrimiento de tendencias y asociaciones, anomalías que puedan ser de interés para procesos clínicos, etc. (Sahoo, 2014). Tomando como ejemplo la información que podemos tener de algunos de los sensores considerados como un pulsioxímetro o un tensiómetro, se puede estudiar la saturación de oxígeno, obteniendo parámetros como el valor promedio, mínimo y máximo a lo largo de un intervalo temporal, el número de caídas en saturación de oxígeno etc. En el caso de la tensión arterial se pueden estudiar los índices de variabilidad y otros parámetros que combinados con la información proveniente de distintas fuentes como los historiales clínicos se refleje de forma fehaciente el estado de salud de una persona y permitir su caracterización a lo largo del tiempo. Una tercera capa de la arquitectura considerada se dedica a la implementación de interfaces de usuario para el acceso a resultados basadas principalmente en HTML5 ofreciendo un sistema (Gachet, 2014) con funciones integradas plenamente operativas para un usuario final.

3.2 Técnicas de Procesamiento del Lenguaje Natural

Las técnicas de Procesamiento del Lenguaje Natural (PLN) se utilizan para mejorar la precisión de los sistemas de información para la categorización de documentos, recuperación de información, sistemas de extracción de conocimiento, generación automática de resúmenes, etc.

Una de las técnicas más utilizadas es la aplicación de ontologías en tareas de PLN para la normalización de términos (que consiste en la transformación de los términos o palabras con el objetivo de reducirlos a formas canónicas que faciliten las correspondencias posteriores en el proceso de búsqueda.) y la organización semántica del contenido en uno o varios planos, que permita indexar el conocimiento poniendo el foco en el contexto de interés del usuario, facilitando por tanto la recuperación inteligente de información relevante para el mismo. Las discrepancias en la terminología utilizada, por diferentes autores, en textos relacionados pueden conllevar una reducción de la efectividad en la detección de la información común. En este sentido, pensamos que la integración de recursos como UMLS (Unified Medical Language System) (Demner-Fushman *et al.* 2010) puede resultar de gran utilidad.

Las técnicas desarrolladas en el ámbito de la minería de textos y de la integración de información se fusionarán a través de una plataforma de código libre que se irá desarrollando paralelamente. Para su elaboración se considerarán las fases clásicas de la Ingeniería del Software, haciendo énfasis en la modularidad y la escalabilidad de la plataforma. Estas fases de realizarán de manera cíclica, siguiendo modelos ágiles de desarrollo en espiral, así como la implementando las funcionalidades en prototipos incrementales que se corresponderán con cada uno de los objetivos parciales de desarrollo del proyecto.

Partiendo de los esquemas básicos de integración de:

a. areas de extracción de información, reconocimiento de entidades y detección de la negación y/o la especulación, la búsqueda de información cross-lingüe y la generación automática de resúmenes en la RI

b. tareas de creación de reglas, clasificación, generación de grafos, agrupamiento sobre datos y la recuperación de imágenes

se definirán esquemas de prototipos centrados en la interfaz, que se irán refinando hasta dar lugar a un conjunto de herramientas con capacidades de minería de texto e integración de información (Romero et al., 2014; L. Borrajo et al., 2015).

3.3 Entornos de evaluación y escenarios

Los escenarios de aplicación de los objetivos del proyecto son tan variados como los intereses de los profesionales de la materia, pero centramos nuestro interés en los siguientes:

- La asistencia a los profesionales sanitarios durante el proceso de toma de

decisiones en el ámbito clínico.
- El acceso a información relevante sobre su estado de salud a pacientes crónicos y dependientes.
- El soporte a la formación basada en la evidencia de los nuevos estudiantes de medicina.

Para dar soluciones a estos escenarios desarrollaremos herramientas que ofrezcan funcionalidades con capacidad para interrelacionar la información de casos clínicos, documentación científica y fuentes de conocimiento específicas, así como herramientas para la visualización de los datos monitorizados a través de los sensores conectados a pacientes. En los tres escenarios que se plantean se analizará toda la información clínica disponible de un paciente para ofrecer al profesional o estudiante información científica que pueda ser de su interés para ese caso concreto.

Los sistemas serán evaluados con usuarios en entornos controlados (hospitalarios, docentes, etc.), con el fin de validar la mejora en las tareas objetivo. Para ello, se definirán tareas concretas de acceso a la información, y se evaluará la capacidad de los usuarios para realizarlas de una manera más efectiva usando dichos sistemas, frente a grupos de usuarios de control que trabajan con sistemas más tradicionales (Villa et al., 2012).

Por otra parte, existen suficientes recursos para la evaluación de la efectividad de muchas tareas de PLN en el ámbito biomédico, muchos de ellos provenientes de evaluaciones competitivas. En los casos que sea posible, se utilizarán dichos recursos para evaluar la efectividad de las nuevas técnicas propuestas (Cruz et al., 2012; Crespo, Mata y Maña, 2013). Finalmente, se realizará una evaluación del sistema completo orientada a medir la usabilidad de la interfaz y la satisfacción de los usuarios.

Bibliografía

Borrajo, L., Seara, A., Iglesias, E.L. 2015 TCBR-HMM: An HMM-based text classifier with a CBR system, Applied Soft Computing, Volume 26, January2015 , Pages 463-473

Crespo, M., Mata, J., Maña, M.J. 2013. Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. *Journal of the American Medical Informatics Association*, 20(6): 1014-1020

Cruz, N.P., Maña, M.J., Mata, J., Pachón, V. 2012. A Machine Learning Approach to Negation and Speculation Detection in Clinical Texts, *Journal of the American Society for Information Science and Technology*, 63(7): 1398- 1410.

Demner-Fushman D, Mork J, Shooshan S, Aronson A. 2010. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. Journal of Biomedical Informatics, 43 (4): 587-594.

Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., Altman, R.B. 2011. Bioinformatics Challenges for Personalized Medicine. Bioinformatics, 27(13):1741-8

Gachet Páez, D., Aparicio, F., de Buenaga, M., y Ascanio, J. R. 2014. Big data and IoT for chronic patients monitoring, UCAmI 2014, pp. 416-423, Belfast, UK

Kumar, D. 2011. The personalised medicine. A paradigm of evidence-based medicine. Ann Ist Super Sanita., 47(1):31-40.

Pandeya S, Voorsluysa W, Niua S, et al. 2012 An autonomic cloud environment for hosting ECG data analysis services. Future Gener Comp Syst 2012;28:147–54

Prajapati, V. 2013. Big data analytics with R and Hadoop. Packt Publishing Ltd.

Romero, R., Seara, A., Iglesias, E.L. and Borrajo, L. 2014. BioClass: A Tool for Biomedical Text Classification. En *Proceedings of the 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, pp. 243-251, Salamanca.

Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, et al. 2014 Heart beats in the cloud: distributed analysis of electrophysiological big data using cloud computing for epilepsy clinical research. J Am Med Inform Assoc 2014. Mar-Apr;21(2):263-71.

Villa, M., Aparicio, F., Maña, M.J. Buenaga, M. 2012. A Learning Support Tool with Clinical Cases Based on Concept Maps and Medical Entity Recognition. En *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pp. 61-70, Lisboa.